ISBN 82–553–0900–4 No.5
April 1994

Confidence intervals from Monte Carlo tests

by

Erik Bølviken and Eva Skovlund

STATISTICAL RESEARCH REPORT – Matematisk institutt, Universitetet i Oslo

Abstract

It is argued that confidence sets can be derived from Monte Carlo tests by exploiting the equivalence between confidence estimation and hypothesis testing. The approach benefits from the wide applicability of this class of tests and the high level accuracy that is passed on to the confidence sets. The main problem is whether the confidence sets are simple enough to be of practical interest. A conservative general approximation is given, but most of the paper deals with exact methods in one-parameter situations. We work with statistics that allow stochastic representations that are almost surely monotone in terms of the parameter of interest. Simulated samples can then be adjusted by varying the parameter and keeping random drawings fixed. By making some selected fractile of such Monte Carlo samples equal to the observed statistic, a critical point with exact confidence level can be determined. A simple theory to compare the Monte Carlo uncertainty to the uncertainty contributed by the data is developed. The main application is to models belonging to the one-parameter exponential class. Other examples considered are location- and scale models, the correlation coefficient, the size parameter in hypergeometric experiments, binomials and rank statistics. The choice of sampling technique is delicate and strongly problem dependent. Exact results, with the exception of pure location-scale models, are rarely possible when there are nuisance parameters, but it is hoped that the basic approach may turn out to be a way to obtain accurate, approximate results under many circumstances.

KEY WORDS: Monte Carlo tests; Exact confidence intervals; Sampling techniques; Exponential class.

1 Introduction

The literature on so-called bootstrap confidence intervals has grown consistently since Efron introduced the idea around 1980. Efron and Tibshirani (1993) and Babu and Rao (1993) have given recent reviews. These methods are able to produce approximate confidence intervals for most statistical models and work both in a parametric and a non-parametric setting. Their formal justification is through asymptotics in the number of observations (Hall, 1988). We shall in this paper suggest an alternative approach that yields, when applicable, exact confidence limits even for small samples. The idea has been inspired by the Monte Carlo tests of Barnard (1963) who pointed out that some selected order statistic from a random sample under the null distribution can be used as critical point. When the (small) Monte Carlo uncertainty is added to the data uncertainty, the resulting test has exact level α for pivotal tests. If estimates of nuisance parameters have been inserted, the level only holds approximately, but the accuracy is good, better than for standard asymptotic tests (Hall and Titterington, 1989).

The problem in applying this technique to confidence estimation is that there are no natural pivots to sample from. One possibility is to utilise the equivalence between hypothesis testing and confidence estimation (Lehmann, 1986). If θ is the parameter of interest, the set of all θ_0 for which the hypothesis H_0:θ=θ_0 is rejected at level α, is a $1-\alpha$ confidence set. For the result to be of practical interest, the confidence set must be simple, an interval say. The question is whether that can be guaranteed with Monte Carlo tests where the rejection regions contain simulation variability created in the computer. Although a general technique to construct approximate *conservative* solutions will be outlined, our present concern is primarily exact methods. One of our points is that exact results in certain practical situations can be obtained by sampling the interest statistic T from a stochastic representation of a particular kind. Consider situations where the distribution of T is fixed by the parameter of interest θ. Suppose there exists a random element Z with distribution not depending on θ so that the pair (θ, Z) yields T through a mapping where θ does not enter either. This means that once a realisation Z^* of Z is available, we have through the mapping a random function $T^* = T^*(\theta)$, the distribution of which coincides with the distribution of T at every θ. We propose to simulate Z a given number of times, *keep this sample fixed*, compute Monte Carlo samples of T from it under variation of θ and keep on until the observed T is matching some prescribed fractile exactly. It will be proved in the next section, through a very simple argument, that this yields exact confidence limits for θ if the mapping $\theta \to T^*$ is monotone for given Z^*. This condition may be restrictive, but there are nevertheless important cases where it can be met.

In applying this idea the focus of research is shifted compared to ordinary bootstrap confidence intervals. Whereas the problem of the bootstrap intervals is to correct bias, the issue now is to find stochastic representations of T possessing the properties demanded, or what amounts to the same thing, designing stochastic algorithms that generate T in the particular way described. Sometimes this is clearcut, as in case of binomials or pure location models, in other instances more tricky. For example, when dealing with conditional

inference in logistic regression, the standard way of simulation does not work, and the day is saved by Markov chain based iterations, as in Gelfand and Smith (1990). In fact nothing prevents us from using complex algorithms with a long string of random variables for the random element Z if only the basic condition cited above is satisfied. This means that there is a battery of general purpose sampling techniques to choose from, see Devroye (1986) and Boswell et al (1993) for general reviews. Neal (1993) gives a bibliography on Markov chain techniques. We shall indicate possible implementations through the discussion of specific examples.

The objective of the paper is to introduce the idea, develop a simple theory to understand the Monte Carlo variability added by the sampling, and make first applications. With one exception only one-parameter models will be considered. The basic technique will be tried out on a number of familiar examples, but the main application is to the one-parameter exponential class where general algorithms are developed. The exact confidence estimates obtained are especially attractive in small-sample situations where the relevance of asymptotics may be unclear, but the usefulness of the approach goes beyond those. Apart from the fact that multi-parameter models can sometimes be reduced to one-parameter ones through conditioning, we do not deal with nuisance parameters. The approach has an obvious potential in this direction, since estimates of unknowns, parameters and distributions alike, can be plugged into the simulator prior to the generation of the Monte Carlo sample. Admittedly, this will destroy the exactness of the methods, but it could prove possible to keep level errors small. The scope for further research is indicated in the closing section.

2 Monte Carlo confidence estimation

Consider a statistic T depending in distribution on a single parameter θ. Assume that T is stochastically increasing in θ, i.e. that the distribution function $F_\theta(t)$ is downwards monotone in θ for any fixed t. Suppose we seek an upper confidence limit for θ based on T. A solution is found by solving

$$F_{\bar{\theta}}(t) = \alpha \tag{2.1}$$

for $\bar{\theta}$. Then $\theta \leq \bar{\theta}$ at confidence coefficient $1 - \alpha$. The level is exact for continuous distributions and approximate and conservative for discrete ones (Lehmann, 1986). The issue we address is how the critical point $\bar{\theta}$ can be determined from simulations when F_θ has an unmanageable analytic expression. This situation occurs frequently in practice, and although the main discussion in the present paper is concerned with the one-parameter case, it is obvious that a given method might also be used after unknown parameters have been estimated.

Several Monte Carlo based confidence intervals are available in the literature. Much has been written about the bootstrapped ones. Garthwaite and Buckland (1992) suggested that the Robbins-Monro algorithm of stochastic approximation can be used to find the

critical point. This certainly works, although it follows from the results in Lai and Robbins (1979) that many simulations are required for accurate determination of $\bar{\theta}$. However, very high accuracy may not necessarily be needed. One might be willing to accept a limited amount of Monte Carlo randomness adding to the uncertainty in the data and tolerate that the same observations do not lead to exactly the same confidence limits under replicated analyses. If so, a more moderate number of simulations would suffice. The point in the Monte Carlo confidence limits proposed, here in analogy to Monte Carlo tests, is to accomodate this philosophy into a precise confidence statement. Most of the discussion will be concerned with one-sided limits. The extension to two-sided ones is obvious.

2.1 The basic construction

Start by recalling the traditional derivation of confidence sets from tests of significance. Let $R(\theta_0)$ be an α rejection region based on T for the hypothesis H_0:$\theta = \theta_0$ and define

$$C = \{\theta | t \notin R(\theta_0)\} \tag{2.2}$$

Since $P_\theta(\theta \in C) = P_\theta(T \notin R(\theta)) = 1 - \alpha$, C is a $1 - \alpha$ confidence set. In the situations we have in mind analytic complexities prevent easy calculation of the rejection regions, and hence of C, but Monte Carlo implementations may be possible. Let $T_1^*(\theta), \ldots, T_m^*(\theta)$ be an independent sample from T when θ is the underlying parameter and write $T_{(1)}^*(\theta) \leq T_{(2)}^*(\theta) \leq \ldots \leq T_{(m)}^*(\theta)$ for the ordered sample. If θ underlies T as well, $T_1^*(\theta), \ldots, T_m^*(\theta), T$ is a $m + 1$-sample from the same parent distribution. Hence, for an atomless distribution,

$$P_\theta(T < T_{(k)}^*(\theta)) = k/(m + 1), \tag{2.3}$$

which identifies a rejection region at exact level $\alpha = k/(m + 1)$ if small values of T are significant. The corresponding version of the confidence set (2.2) is

$$C = \{\theta | t \geq T_{(k)}^*(\theta)\}, \tag{2.4}$$

with confidence level $1-\alpha = 1-k/(m+1)$. Although C is an exact confidence set whatever method is used to generate the Monte Carlo sample, the procedure must in practice lead to some simple set, preferably an interval. An approximate solution is always available. Define

$$\bar{\theta} = \inf\{\theta | T_{(k)}^*(\theta) > t\}. \tag{2.5}$$

Then $C \subset (-\infty, \bar{\theta})$ and since the level of C was exact

$$P_\theta(\theta < \bar{\theta}) \geq 1 - k/(m + 1), \tag{2.6}$$

so that $\bar{\theta}$ is a conservative, upper confidence point for θ.

The question is how much the real confidence level deviates from its lower bound. It is possible to arrange things so that the *same* random drawings go into the Monte Carlo sample for *different* θ. This means that $T_{(k)}^*(\theta)$ in (2.5) traces out a curve that is smooth

in θ, except possibly for a few jumps, and, by the stochastic monotonicity, drifting upwards as θ is moved from small to large values, although there may be local bumps. It is *these perturbations from strict monotonicity that destroy the exact confidence statements.* The issue is the same for discrete distributions. There are reasons to hope that the level inaccuracy is not too great. We shall comment on this in the closing section. The rest of the paper deals with situations permitting exact methods.

2.2 Exact confidence limits

Henceforth assume that the statistic T allows a stochastic representation of the form

$$T = h(\theta, Z), \tag{2.7}$$

where Z is some random vector, *with distribution not depending on* θ and h is some known function, which is upwards monotone in θ for given z. In theory, any atomless statistic T that is stochastically increasing in θ can be represented in this way. Simply take $T = F_\theta^{-1}(U)$ for a uniform random number on $(0,1)$. However, this is pointless, since if F_θ^{-1} is so simple that T can be sampled by inversion, an exact solution to (2.1) would have been available in the first place. What we are assuming is a representation that is practical to sample from.

Let $Z_1^*, ..., Z_m^*$ be a sample from Z. Define

$$T_j^*(\theta) = h(\theta, Z_j^*), \qquad j = 1, ..., m. \tag{2.8}$$

so that $T_1^*(\theta), ..., T_m^*(\theta)$ is a sample from T when θ is the underlying parameter. Keep the Z-sample fixed and vary θ. The conditions assumed mean that each sampled $T_j^*(\theta)$ is upwards monotone in θ. This extends to the order statistic $T_{(k)}^*(\theta)$ in (2.3) so that the exact confidence set $C = (-\infty, \bar{\theta})$, where $\bar{\theta}$ was defined in (2.5). (2.6) is now valid as an equality. If $T_{(k)}^*(\theta)$ is continuous in θ, then $\bar{\theta}$ is the solution of

$$T_{(k)}^*(\bar{\theta}) = t, \tag{2.9}$$

but the condition that the sampled variables are continuous functions in θ is not so obvious as it may seem. For example, the algorithm in 5.3 below does not possess this property even if the distribution function F_θ of the observed statistic does.

The actual solving of equation (2.9) or (2.5) requires a little elaboration. We have used the numerically safe bisection method (Press et al, 1986) which is very easy to implement. First locate θ-values θ_1 and θ_2 on each side of $\bar{\theta}$ by checking the sign of $T_{(k)}^*(\theta_j) - t$, $j = 1, 2$. Successive halvings of the interval (θ_1, θ_2) after examining the sign of $T_{(k)}^*(\theta) - t$ at the mid-points, will set up a numerical iteration that is certain to converge to a solution of (2.9) if one exists.

2.3 Ties

The discussion in 2.2 neglected ties. If T has a discrete distribution, and θ is a continuously varying parameter, (2.9) has an interval on the real line as solution set. The right end point corresponds to the conservative bound (2.5).

It is perhaps more reasonable to seek an exact treatment of ties in the present context than elsewhere in statistics, since there is already a Monte Carlo element present. The simplest approach feeds on the way randomisation is handled in the traditional theory of confidence estimation (Lehmann, 1986). Suppose for simplicity that T is integer-valued. Replace T by $\tilde{T} = T + U$, where U is a uniform random number on (0,1). By (2.7)

$$\tilde{T} = h(\theta, Z) + U, \tag{2.10}$$

and the monotonicity condition in 2.2 holds for the now continuously valued $\tilde{T}$. Exact confidence points for θ therefore follow by comparing the 'observed' $\tilde{T}$, say $\tilde{t} = t + u$, to a Monte Carlo sample $\tilde{T}^*_j(\theta) = T^*_j(\theta) + U^*_j$, $j = 1, \ldots, m$, solving (2.5) or (2.9) as before.

An alternative (and equivalent) way of handling ties due to Jöckel (1986) may be more transparent. Let

$$\begin{aligned} k^-(\theta) &= \sup\{j | T^*_{(k-j)}(\theta) = T^*_{(k)}(\theta)\} \\ k^+(\theta) &= \sup\{j | T^*_{(k+j)}(\theta) = T^*_{(k)}(\theta)\}. \end{aligned} \tag{2.11}$$

The sum $k^-(\theta) + k^+(\theta) + 1$ is then the number of simulations tying with $T^*_{(k)}(\theta)$. Introduce

$$q^*_k(\theta) = \frac{k^-(\theta) + 1}{k^-(\theta) + k^+(\theta) + 2} \tag{2.12}$$

and fix some uniform random number U^*. Define

$$H^*(\theta) = \begin{cases} T^*_{(k)}(\theta) - t & \text{if } T^*_{(k)}(\theta) \neq t \\ q^*_k(\theta) - U^* & \text{if } T^*_{(k)}(\theta) = t. \end{cases} \tag{2.13}$$

Modify the critical point $\bar{\theta}$ to

$$\bar{\theta} = \inf\{\theta | H^*(\theta) > 0\}, \tag{2.14}$$

which is now exact at level $k/(m+1)$. To prove this, treat $Z^*_1, \ldots, Z^*_m$ and U^* as given. It is not difficult to deduce from the monotonicity of $T^*_{(k)}(\theta)$ that if $H^*(\theta) > 0$, then also $H^*(\theta_1) > 0$ for any $\theta_1 \geq \theta$. This means that $\theta \geq \bar{\theta}$ if and only if $H^*(\theta) > 0$. (Note that $H^*(\theta) \neq 0$ with probability one.) Hence

$$\begin{aligned} P_\theta(\theta \geq \bar{\theta}) &= P_\theta(H^*(\theta) > 0) \\ &= P_\theta(T^*_{(k)}(\theta) > T) + P_\theta(T^*_{(k)}(\theta) = T,\ U^* \leq q^*_k(\theta)). \end{aligned}$$

It follows from Jöckel (1986), see Proposition 2.1 in that paper, that the last line equals $k/(m+1)$.

3 The Monte Carlo variability

3.1 Notation and results

The impact of the Monte Carlo randomness can be analysed theoretically under the conditions in 2.2. A more elaborate mathematical notation is then needed. Let $\bar{\theta}_m(\alpha)$ (up to now only $\bar{\theta}$) be the upper critical point if m simulations are used and the confidence coefficient is $1-\alpha$. Write $\bar{\theta}(\alpha)$ for the corresponding solution of (2.1). Similarly, let $\underline{\theta}_m(\alpha)$ and $\underline{\theta}(\alpha)$ be lower critical points and

$$L_m(\alpha) = \bar{\theta}_m(\alpha) - \underline{\theta}_m(\alpha) \tag{3.1}$$

the length of an ordinary two-sided, equally tailed, Monte Carlo $1-2\alpha$ confidence interval with $L(\alpha) = \bar{\theta}(\alpha) - \underline{\theta}(\alpha)$ holding the same meaning for the non-random interval. Throughout this section overline refers to upper critical points, underline to lower ones and symbols with neither to interval length. The subscript m, as in $\underline{\theta}_m(\alpha)$ signifies a critical point based on m simulations whereas no subscript, for instance $L(\alpha)$ stands for a confidence interval or point calculated from the exact distribution.

Introduce

$$\bar{G}_m(x;\alpha) = P(\bar{\theta}_m(\alpha) \leq x), \tag{3.2}$$

and similarly $\bar{G}(x;\alpha)$ for the distribution function of $\bar{\theta}(\alpha)$. Their dependence on the underlying parameter θ is suppressed in the notation. There is a simple connection between $\bar{G}_m$ and $\bar{G}$. Let B be a Beta-distributed random variable with parameters $\alpha(m+1)$ and $(1-\alpha)(m+1)$. Then

$$\bar{G}_m(x;\alpha) = E^*\{\bar{G}(x;B)\}, \qquad \text{(upper critical limits)} \tag{3.3}$$

where the expectation E^* on the right is taken with repect to the random variable B. The proof is given below. There are analogous counterparts for the lower critical point and for interval length, for example

$$G_m(x;\alpha) = E^*\{G(x;B)\} \qquad \text{(interval length)}. \tag{3.4}$$

These results clarify how the Monte Carlo based methods are connected to the corresponding non-random versions. We are evidently working with *random confidence coefficients* that have been drawn, *prior to the analysis*, from a distribution that has been fixed by the number of simulations selected. Note that B follows the same distribution whatever model used. It is illuminating to look at its properties. The mean is α and the variance $\alpha(1-\alpha)/(m+2)$. Thus the distribution becomes more and more concentrated around α as m grows. In the limit, as $m \to \infty$, $\bar{\theta}_m(\alpha) \to \bar{\theta}(\alpha)$ and $L_m(\alpha) \to L(\alpha)$ under some weak assumption of continuity with respect to α. A similar result on Monte Carlo tests is given by Jöckel (1986).

We have tabulated the percentiles of this basic Beta-distribution in Table 1. It seems

		Percentiles				
m	α	0.05	0.25	0.50	0.75	0.95
99	0.005	0.0000	0.0005	0.0023	0.0066	0.0192
	0.025	0.0058	0.0135	0.0219	0.0332	0.0548
	0.050	0.0201	0.0341	0.0470	0.0627	0.0901
999	0.005	0.0020	0.0034	0.0047	0.0063	0.0091
	0.025	0.0175	0.0215	0.0247	0.0281	0.0336
	0.050	0.0392	0.0452	0.0497	0.0545	0.0618
9999	0.005	0.0039	0.0045	0.0050	0.0055	0.0062
	0.025	0.0225	0.0239	0.0250	0.0260	0.0276
	0.050	0.0465	0.0485	0.0500	0.0515	0.0536

Table 1: Percentiles of the Beta-distribution for $\alpha = 0.005$, 0.025, and 0.05 under variation of m. (Note that the α-values correspond to 0.99, 0.95, and 0.90 confidence intervals.)

as if $m = 99$ corresponds to an uncomfortably large spread in confidence levels, whereas $m = 999$ is acceptable. The importance of such Monte Carlo uncertainty can be studied through Pitman efficiencies, as Jöckel (1986) did for Monte Carlo tests, but this breaks with the small sample motivation of the paper, and we shall leave that approach aside. Alternatively, the mixing formulas (3.3) and (3.4) can be used to compare means and variances of the confidence estimates. The results are identical for one-sided and two-sided intervals and will be given only for the former.

Let

$$\bar{\mu}_m(\alpha) = E\{\bar{\theta}_m(\alpha)\} \tag{3.5}$$

$$\bar{\tau}_m^2(\alpha) = \mathrm{var}\{\bar{\theta}_m(\alpha)\} \tag{3.6}$$

be the *unconditional* mean and variance of the Monte Carlo critical points, i.e. encompassing both data and simulation variability. (We use the convention that when E and var are not *-marked, both types of uncertainties are included. By contrast, the *-marked versions are taken with respect to the Monte Carlo randomness conditional on the data.) Moreover, let $\bar{\mu}(\alpha)$ and $\bar{\tau}^2(\alpha)$ be the analogous quantities for the non-random critical limits. Then

$$\bar{\mu}_m(\alpha) = E^*\{\bar{\mu}(B)\} \tag{3.7}$$

$$\bar{\tau}_m^2(\alpha) = E^*\{\bar{\tau}^2(B)\} + \mathrm{var}^*\{\bar{\mu}(B)\}. \tag{3.8}$$

Equations (3.7) and (3.8) can be deduced from (3.3). Condition $\bar{\theta}_m(\alpha)$ on B and use the rule of double expectation for (3.7) and the analogous double variance formula for (3.8).

3.2 Proofs

Equations (3.3) and (3.4) must be verified. Suppose first that the distribution of T has no atoms. Consider (3.3). Recall that $T^*_{(k)}(x)$ increases with x. Hence, by (2.5), the events $\bar{\theta}_m(\alpha) \leq x$ and $t \leq T^*_{(k)}(x)$ occur simultaneously, except, possibly on a set of measure null relating to the boundaries. Thus,

$$\begin{aligned} P_\theta\{\bar{\theta}_m(\alpha) \leq x\} &= P_\theta\{T \leq T^*_{(k)}(x)\} \\ &= P_\theta\{F_x(T) \leq F_x(T^*_{(k)}(x))\} \\ &= P_\theta\{F_x(T) \leq B\}, \end{aligned} \tag{3.9}$$

where $B = F_x\{T^*_{(k)}(x)\}$ applies F_x to the k'th order statistic of the Monte Carlo sample $T^*_1(x), ..., T^*_m(x)$. Since F_x is monotone, nothing is changed if F_x is used first and the ranking done afterwards. Thus $B = U^*_{(k)}$, where $U^*_j = F_x\{T^*_j(x)\}$, $j = 1, ..., m$. This is a sample of uniforms on $(0, 1)$. The k'th order statistic from such samples is known to be Beta-distributed with parameters k and $m + 1 - k$, which on inserting $\alpha(m + 1)$ for k become $\alpha(m + 1)$ and $(1 - \alpha)(m + 1)$. In other words, B has the distribution asserted and is also stochastically independent of T, since its variability comes from the Monte Carlo experiment. The last line in (3.9) can be rewritten

$$\bar{G}_m(x; \alpha) = P_\theta\{F_x(T) \leq B\}. \tag{3.10}$$

Fix $B = b$ and perform the steps in (3.9) in reverse. Then

$$\begin{aligned} P_\theta(F_x(T) \leq b) &= P_\theta\{T \leq F_x^{-1}(b)\} \\ &= P_\theta\{\bar{\theta}(b) \leq x\} \\ &= G(x; b), \end{aligned}$$

which in combination with (3.10) yields (3.3), since B and T are independent.

The analogous result (3.4) for the length of a two-sided Monte Carlo interval is true even if the lower and upper critical point are taken from the *same* round of simulations so that there is stochastic dependence in their joint simulation variability. Write

$$G_m(\underline{x}, \bar{x}; \alpha) = P_\theta\{\underline{\theta}_m(\alpha) \leq \underline{x}, \bar{\theta}_m(\alpha) \leq \bar{x}\}$$

for the joint distribution function of $(\underline{\theta}_m(\alpha), \bar{\theta}_m(\alpha))$. It can be proved through exactly the same steps that lead to (3.3) that

$$G_m(\underline{x}, \bar{x}; \alpha) = E^*\{G(\underline{x}, \bar{x}; B)\}, \tag{3.11}$$

where B is the same mixing variable as before. (3.4) is a consequence of this extension of (3.3).

If the distribution of T is not continuous (say integer valued), the results still stand, but now as relationships in terms of methods based on $\tilde{T}$, as explained in 2.3. This means that a Monte Carlo confidence point or interval that incorporates exact treatment of ties is connected through (3.3) and (3.4) to a version based on T that employs randomisation in the traditional way.

4 Simple examples

Monte Carlo confidence intervals are in this section derived in a number of familiar examples. The procedures for the location and scale models can be found in Ripley (1987), but the rest are believed to be new. Many of the methods are strikingly simple, and several of them may be useful for simultaneous inference, as in Beran (1988). The limiting, non-random counterparts, as the number of simulations $m \to \infty$, are in some cases included in standard software packages, in others not in common use. The formulae for the mean and variance of the confidence points become particularly simple for models of location and scale and are used to gain insight into the importance of simulation uncertainty in such models. It is tempting to extrapolate to other situations. Ties are largely disregarded.

4.1 Pure location

The simplest example conceivable is the pure location model

$$T = \theta + \sigma\varepsilon, \tag{4.1}$$

where ε has a known distribution with unit variance and σ is known. (Note that σ is used in a meaning different from the usual one in that it refers to the standard deviation of T rather than the individual observations). Clearly $T^*_{(k)}(\theta) = \theta + \sigma\varepsilon^*_{(k)}$, where $\varepsilon^*_{(k)}$ is the k'th order statistic of an ε Monte Carlo sample. Hence

$$\bar{\theta}_m(\alpha) = t - \sigma\varepsilon^*_{(k)} \tag{4.2}$$

is an (exact) upper confidence limit for θ.

The consequences of running only a finite number of simulations is easy to understand in this situation. Let $c(\alpha)$ be the lower α-percentile of T when $\sigma = 1$. Simple calculations using (3.7) and (3.8) yield

$$\bar{\mu}_m(\alpha) = \bar{\mu}(\alpha) + \sigma a_m$$
$$\bar{\tau}_m(\alpha) = \bar{\tau}(\alpha) b_m,$$

where

$$a_m = c(\alpha) - E^*\{c(B)\}$$
$$b_m^2 = 1 + \mathrm{var}^*\{c(B)\}.$$

If m simulations are run, the Monte Carlo critical point will, on average, be placed an amount σa_m above the non-random counterpart whereas its standard deviation is inflated by the factor b_m. The coefficients a_m and b_m are estimated for different distributions in Table 2 under variation of m and α. When recalling that confidence limits are generally a couple of standard deviation units above the point estimate, it can surely be concluded that the (extra) effect of the Monte Carlo is small compared to the uncertainty in the data. At low confidence levels there are clear signs, primarily in a_m defining bias, that $m = 99$ repetitions are not enough.

		Normal		t_3		Gamma			
						Lower		Upper	
m	α	a_m	b_m	a_m	b_m	a_m	b_m	a_m	b_m
99	0.005	0.353	1.1849	-	-	0.036	1.0040	0.864	1.8036
	0.025	0.072	1.0392	0.218	1.0194	0.014	1.0037	0.153	1.1265
	0.050	0.038	1.0232	0.080	1.0479	0.009	1.0036	0.074	1.0629
999	0.005	0.031	1.0126	0.170	1.1747	0.004	1.0005	0.070	1.0546
	0.025	0.007	1.0036	0.020	1.0115	0.001	1.0004	0.014	1.0114
	0.050	0.003	1.0022	0.007	1.0039	0.001	1.0003	0.007	1.0059
9999	0.005	0.003	1.0012	0.016	1.0145	0.000	1.0000	0.007	1.0051
	0.025	0.000	1.0003	0.002	1.0011	0.000	1.0000	0.002	1.0011
	0.050	0.000	1.0002	0.001	1.0004	0.000	1.0000	0.001	1.0006

Table 2: The coefficients a_m and b_m for the pure location model under variation of m and α. The computations were from 50000 simulations of B using the BLSS package. (Results left out were due to numerical trouble).

4.2 Pure scale

Equally transparent results are obtained for pure scale where the interest statistic S is of the form

$$S = \theta Z, \tag{4.3}$$

where Z is a positive random variable with mean 1. In this case

$$\bar{\theta}_m(\alpha) = \frac{s}{Z^*_{(k)}}. \tag{4.4}$$

The limiting, non-random point as $m \to \infty$ is $\bar{\theta}(\alpha) = s/c(\alpha)$ where $c(\alpha)$ is the lower α percentile of Z. In this case (3.7) and (3.8) become

$$\bar{\mu}_m(\alpha) = \bar{\mu}(\alpha)a_m$$
$$\bar{\tau}_m(\alpha) = \tau(\alpha)b_m$$

where

$$a_m = E^*\{\frac{c(\alpha)}{c(B)}\}$$
$$b_m^2 = E^*\{\frac{c(\alpha)}{c(B)}\}^2 + \{\mathrm{var}(Z)\}^{-1}\mathrm{var}^*\{\frac{c(\alpha)}{c(B)}\}.$$

The impact of the Monte Carlo variability can be understood through the constants a_m and b_m Note, in particular, that b_m goes up with the inverse of var(Z). This behaviour is quite different from the location case and to control the *relative* importance of the Monte Carlo uncertainty, the number of simulations m must be linked to the number of observations underlying the statistic T. Table 3, where a_m and b_m have been computed

		χ^2 distribution							
		$\nu = 5$				$\nu = 20$			
		Upper		Lower		Upper		Lower	
m	α	a_m	b_m	a_m	b_m	a_m	b_m	a_m	b_m
99	0.005	-	-	0.8526	3.048	1.135	5.988	0.9114	7.325
	0.025	1.044	1.394	0.9620	2.200	1.025	3.585	0.9789	5.041
	0.050	1.020	1.320	0.9785	1.945	1.012	3.234	0.9886	4.374
999	0.005	1.024	1.130	0.9863	1.910	1.012	2.511	0.9919	4.017
	0.025	1.005	1.096	0.9963	1.493	1.003	2.068	0.9980	2.936
	0.050	1.002	1.092	0.9979	1.374	1.001	1.949	0.9989	2.587
9999	0.005	1.003	1.033	0.9986	1.356	1.001	1.601	0.9992	2.423
	0.025	1.000	1.029	0.9996	1.179	1.000	1.420	0.9998	1.843
	0.050	1.000	1.030	0.9998	1.132	1.000	1.372	0.9999	1.672

Table 3: The coefficients a_m and b_m for the pure scale model under variation of α and m. χ^2-distributions with $\nu = 5$ and $\nu = 20$ were used as examples and both lower and upper limits are considered. The computations were from the same simulations of B as in Table 2.

for two χ^2-distributions, illustrates the effect of the Monte Carlo randomness. Comparison with Table 2 suggests that m should now be taken somewhat larger. Note that b_m, defining standard deviation relative to the non-random confidence point, is much higher for $\nu = 20$.

4.3 Location-scale models

Models that are of the pure location-scale type, are exceptional in that exact methods are available even when there is a nuisance parameter. Suppose the interest statistic $T = \theta + \sigma\varepsilon$ is of the same type as in 4.1 above, but now with unknown σ. Let $S = \sigma Z$ be an estimate of σ of the form studied in 4.2. Assume that (ε, Z) has a known distribution not depending on (θ, σ). There is no need to require ε and Z to be independent. Many problems of practical interest can be cast in this form.

To construct exact Monte Carlo confidence bounds of the t-type, generate m sampled pairs (ε_j^*, Z_j^*), $j = 1, \ldots, m$ and let

$$\bar{\theta} = t - s(\frac{\varepsilon^*}{Z^*})_{(k)} \tag{4.5}$$

where $(\varepsilon^*/Z^*)_{(k)}$ is the k'th order statistic of all "t-ratios" $\varepsilon_1^*/Z_1^*, \ldots, \varepsilon_m^*/Z_m^*$. (4.5) is now an exact upper critical point for θ. To see this observe that

$$\begin{aligned} P_{\theta,\sigma}(\theta \geq \bar{\theta}) &= P_{\theta,\sigma}(\theta \geq T - S(\frac{\varepsilon^*}{Z^*})_{(k)}) \\ &= P_{\theta,\sigma}(\frac{T-\theta}{S} \leq (\frac{\varepsilon^*}{Z^*})_{(k)}) \end{aligned}$$

$$= P(\frac{\varepsilon}{Z} \leq (\frac{\varepsilon^*}{Z^*})_{(k)}).$$

Note that both θ and σ have vanished due to the special nature of the model. As before the confidence coefficient becomes $k/(m+1)$.

Formulae similar to those derived in 4.1 for the mean and standard deviation of the critical point can be found. Although the theory in section 3 does not quite cover the present situation, the results given there may easily be extended. The expressions for $\bar{\mu}_m(\alpha)$, $\bar{\tau}_m(\alpha)$ and a_m remain about the same, but b_m is different. Let $\text{var}(\varepsilon) = 1$ as in 4.1. If a possible correlation between ε and Z is ignored (it can easily be accomodated at the expense of a longer formula), it turns out that

$$b_m^2 = \frac{1 + (EZ)^2 \text{var}^*\{c(B)\} + \text{var}(Z) E^*\{c(B)^2\}}{1 + \text{var}(Z) c^2(\alpha)}.$$

When $E(Z) \to 1$ and $\text{var}(Z) \to 0$ so that the bias and random error in the σ-estimate vanishes, the formula collapses to the one in 4.1.

4.4 The correlation coefficient

Exact interval estimation is possible for a correlation coefficient ρ under normal, or more generally, elliptically contoured distributions. The point is to utilise that the second variable of a pair (X, Y) in such models has the representation

$$Y = \rho X + (1 - \rho^2)^{1/2} Z, \tag{4.6}$$

where $\text{corr}(X, Z) = 0$ and (X, Z) belongs to the same elliptic family.

Let $(X_1, Y_1), ..., (X_n, Y_n)$ be a sample distributed as (X, Y) and let $Z_1, ..., Z_n$ be the analogy to Z in (4.6). Straightforward algebra shows that the sample correlation R between X's and Y's may be written

$$\frac{R}{(1 - R^2)^{1/2}} = \{\frac{\rho}{(1 - \rho^2)^{1/2}} \frac{S_X}{S_Z} + R_{XZ}\}\{1 - R_{XZ}^2\}^{-1/2}, \tag{4.7}$$

where S_X and S_Z are the standard deviations for $X_1, ..., X_n$ and $Z_1, ..., Z_n$ respectively and R_{XZ} their correlation coefficient. It follows that R is an increasing function of ρ for given X's and Z's, precisely as demanded for exact inference. Hence, we may draw X- and Z-samples, store them and compute Monte Carlo versions, $R_1^*(\rho), ..., R_m^*(\rho)$, by (4.7), under variation of ρ. An exact upper confidence point is then found by solving

$$R_{(k)}^*(\bar{\rho}) = r \tag{4.8}$$

where r is the data value of the observed correlation.

4.5 Population size from hypergeometric experiments

Consider n_2 units sampled from a population of unknown size N. Let X be the number of units of a particular type in the sample. We address the issue of setting confidence limits to N when the number of units of the given type in the population is a known number n_1. This problem arises in capture-recapture experiments where n_1 is the number of animals marked after a first catch and x the number of marked animals subsequently caught. Garthwaite and Buckland (1992) used stochastic approximation to obtain interval estimates of N. We shall now use the technique of the present paper to construct exact intervals (although it might in the present instance be possible to solve (2.1) directly). The following algorithm is a consequence of our basic procedure.

Proceed conditionally on n_1 and n_2. The issue is how to generate the sampled versions $X^*(N)$ of X. Store n_2 uniform random numbers $U_1^*, \ldots, U_{n_2}^*$ in the computer. Define recursively

$$X_j^*(N) = X_{j-1}^*(N) + I(U_{j-1}^* \leq \frac{n_1 - X_{j-1}^*(N)}{N-(j-1)}, \qquad j = 1, ..., n_2, \tag{4.9}$$

where $I(A) = 1$ if event A is true and 0 otherwise. Start the recursion at $X_0^*(N) = 0$. Then $X^*(N) = X_{n_2}^*(N)$ is a simulation of X based on population size N. It is easy to check that $X^*(N)$ is downwards monotone in N, when $U_1^*, ..., U_{n_2}^*$ are fixed. An upper confidence point of N is the solution of

$$X_{(m+1-k)}^*(\bar{N}) = x. \tag{4.10}$$

Note that the upper rather than the lower fractile of the Monte Carlo sample must be used, since the monotonicity is reversed from the other examples. (4.10) gives an approximate $k/(m+1)$ critical point given n_1 and n_2 and hence an unconditional critical point as well. The procedure is inexact due to the possibility of ties. It could have been made exact by employing randomisation.

4.6 Binomial p

The number of successes in a Bernoulli trial can be represented as

$$X = \sum_{i=1}^{n} I(U_i \leq p), \tag{4.11}$$

where I is an indicator function as before, and $U_1, U_2, \ldots, U_n$ are uniform random numbers. Note that the summands, and hence X, can only go up when p is raised.

To set an upper confidence limit to p generate m sets of n uniform random numbers, U_{ji}^*, $i = 1, \ldots, n$ for set j. Define simulated binomials $X_j^*(p)$ from (4.11). Ignore ties. An upper confidence point $\bar{p}$ is to be determined by adjusting p so that $X_{(k)}^*(p)$ equals the observed x. Note that

$X_j^*(p) > x$ if and only if $U_{j(x+1)}^* \leq p$

where $U^*_{j(1)} \leq \ldots \leq U^*_{j(n)}$ are the order statistics for set j. Hence, if $B^*_j(x) = U^*_{j(x)}$, then $X^*_{(k)}(p) > x$ is equivalent to $B^*_{(k)}(x+1) \leq p$ and

$$\bar{p} = B^*_{(k)}(x+1) \tag{4.12}$$

becomes a confidence bound, slightly inexact due to the negligence of ties. $B^*_j(x+1)$ is Beta-distributed with parameters $x+1$ and n. The Monte Carlo critical point at level $1 - k/(m+1)$ can therefore be found be sampling m such Beta-distributed random variables and taking the k'th smallest as $\bar{p}$.

Randomisation is required (even in the non-random case) for exact adjustment to the desired confidence level. It follows from the argument above that $X^*_{(k)}(p) = x$ if and only if $B^*_{(k)}(x) \leq p < B^*_{(k)}(x+1)$. The point corresponding to the exact level is to be selected within this interval. This must be done numerically as outlined in 2.3.

4.7 Ranks

Another potential area of applications is rank based procedures. Consider as a specific example the problem of setting confidence limits to a shift parameter δ in a two sample situation. One of the standard procedures (see e.g. p. 72 in Hollander and Wolfe, 1973) is to solve (2.1) for the Wilcoxon statistic, which leads to a distribution free upper confidence point. The Monte Carlo analogue to be developed may have practical significance for simultaneous confidence intervals. The derivation is similar to the binomial case, and an explicit formula for the confidence point $\bar{\delta}$ can be found if ties are ignored.

Work from the Mann-Whitney version, denoted X, which has the following representation. Let $Z_1, \ldots, Z_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be two samples from some common distribution. Then

$$X = \sup\{i | D_{(i)} \leq \delta\}, \tag{4.13}$$

where $D_{(1)} \leq D_{(2)} \leq \ldots \leq D_{(n_1 n_2)}$ are the ordered differences $Z_i - Y_l$. When dealing with simulations, $X^*(\delta)$ and $D^*_{(i)}$ in our usual notation, we clearly have

$$X^*(\delta) > x \text{ if and only if } D^*_{(x+1)} \leq \delta.$$

The rest of the argument now follows the preceding subsection exactly. If $V^*_j = D^*_{(x+1)}$ for replication j, then

$$\bar{\delta} = V^*_{(k)} \tag{4.14}$$

is the critical upper limit for the shift parameter δ. The whole operation is distribution-free, and we may use any model of our choice for the generation of the Monte Carlo samples.

5 The one parameter exponential class

Consider statistics T with densities

$$f_\beta(t) = c(\beta)\exp(\beta t)f_0(t). \tag{5.1}$$

Many multiparameter models can be reduced to this form through conditioning. T is stochastically increasing in β and is thus a candidate for the general technique of the paper. Note that exact confidence limits would be available by numerical methods if it were practical to compute the reference density f_0. That is an assumption we are *not* going to make. Poisson and logistic regression with one covariate illustrate our point. The sufficient statistic for the slope is given in (5.3) below. Its distribution is not a simple one under either model, although algorithms have been developed in particular cases for conditional logistic regression, see Vollset et al (1991) and the references therein. By contrast, sampling is straightforward, though not necessarily when the special demands from 2.2 are imposed. We first explore the difficulties through the two examples mentioned and run into trouble with one of them. Two different *general* algorithms that work are presented next, and then numerical illustrations.

5.1 Poisson and logistic regression

Consider poisson and logistic regression with one covariate. In standard GLIM notation take $\eta = \beta_0 + \beta x$ and, assuming canonical links, $\log(\lambda) = \eta$ for poisson and $\text{logit}(p) = \eta$ for the logistic. Suppose there are covariates $x_1, \ldots, x_n$ on n units and let $Y_1, \ldots, Y_n$ be the observations, either poisson or binary. In either case, conditioning with respect to

$$S = \sum_{i=1}^{n} Y_i \tag{5.2}$$

reduces the densities of the interest statistic

$$T = \sum_{i=1}^{n} x_i Y_i \tag{5.3}$$

to the form (5.1), the intercept parameter β_0 disappearing. T is to be simulated in the special way described in section 2, i.e. when the random drawings are held fixed, the sampled value $T^* = T^*(\beta)$ is to go up with β.

Start with poisson regression. The density of $(Y_1, \ldots, Y_n)$ given $S = s$ is that of a multinomial based on s trials and success probabilities

$$q_i = \frac{\exp(\beta x_i)}{\sum_{i'} \exp(\beta x_{i'})}. \tag{5.4}$$

To construct a sampled version of T, order $x_1, \ldots, x_n$ according to decreasing size, i.e. imagine that $x_1 \geq x_2 \geq \ldots \geq x_n$. Then draw s uniform random numbers $U_1^*, \ldots, U_s^*$, and define for $j = 1, \ldots, s$

$$i^*(j) = \inf\{i | q_1 + \cdots + q_i \geq U_j^*\}. \tag{5.5}$$

The integer $i^*(j)$ at trial j has then been selected within the set $\{1, 2, \ldots, n\}$ with probability q_i. This is indeed one of the standard ways of simulating multinomial trials. More efficient procedures can be found in Devroye (1986), but they do not lead to algorithms satisfying our basic condition. Take

$$T^*(\beta) = \sum_{j=1}^{s} x_{i^*(j)}, \tag{5.6}$$

as the sampled version of T. Then $T^*(\beta) \leq T^*(\beta_1)$ for $\beta \leq \beta_1$ if the uniforms $U_1^*, ..., U_s^*$ are held fixed. To see this, note that (5.5) for fixed U_j^* returns an integer $i^*(j)$ that is not greater under β_1 than under β. With the particular ordering of $x_1, ..., x_n$ the contribution $x_{i^*(j)}$ to the sum (5.6) is therefore made at least as large under β_1 as under β.

The monotonicity required for exact Monte Carlo confidence intervals is thus available for poisson regression, but the story is different for logistic regression. Although the very same technique can be used to generate samples of T, there is one notable difference. The integers $i^*(j)$ are now sampled *without replacement.* This means that the preceding closing argument that lead to the algebraic monotonicity in β is no longer valid. Indeed, it is easy to see from a simple example ($s = 3$ is sufficient) that the natural sequential way of sampling T does not work. An approximate (and conservative) procedure could still be found along the lines suggested in section 2, but exact solutions are available by other means.

5.2 Importance sampling

The algorithm in both this and the next subsection is based on our ability to sample from densities of the form (5.1). Select some paremeter of reference, say $\beta = \delta$, from considerations of sampling effciency and start by drawing a sample $t_1, \ldots, t_\nu$ from f_δ. As in 5.1 order according to decreasing size so that $t_1 \geq t_2 \geq \ldots \geq t_\nu$ and store the sequence in the computer. Compute

$$q_i = \frac{\exp\{(\beta - \delta)t_i\}}{\sum_{i'} \exp\{(\beta - \delta)t_{i'}\}}, \qquad i = 1, \ldots, \nu, \tag{5.7}$$

and sample an integer i^* from this distribution in the way described for poisson regression, i.e. by (5.5). Return

$$T^*(\beta) = t_{i^*}. \tag{5.8}$$

As in 5.1, the ordering of $t_1, ..., t_\nu$ ensures monotonicity in terms of β.

The second step may be repeated. Suppose m drawings are taken (with replacement) from the pool $t_1, \ldots, t_\nu$ created at the first step. It is proved in 5.6 below that a sample from (5.1) appears in the limit as $\nu \to \infty$. The algorithm leads to exact confidence bounds for infinitely large ν, since the strict monotonicity condition in 2.2 is satisfied. It is possible to generate the candidates from more general densities of the form $cf_0(t)\exp\{w(t)\}$ for some function w, provided $(\beta - \delta)t_j$ in (5.7) is replaced by $\beta t_j - w(t_j)$.

5.3 Sampling by Markov chain iterations

Besag and Clifford (1989) who were concerned with hypothesis testing, combined exact Monte Carlo inference with Markov chain based sampling in a particularly elegant way. They noted that to deduce (2.3) it is enough that the sample $T, T_1^*(\theta), \ldots, T_m^*(\theta)$ is exchangeable. It is not immediately apparent how this observation can be utilised. However, Besag and Clifford invented an ingenious trick which we first present in a general setting before reverting to the one-parameter exponential class.

Let the parameter of interest be β rather than θ and let $q_\beta(t'|t)$ be the transition probabilities of a reversible, irreducible Markov chain on the space of the test statistic. We shall introduce a specific construction for q_β later, but for the moment assume that the density f_β is the equilibrium distribution of the Markov chain. This means, in particular, that if the Markov process is started as the observed t, which is known to have been generated from f_β, then all subsequent simulations from the process will possess this distribution too. Proceed as set down in Figure 1.

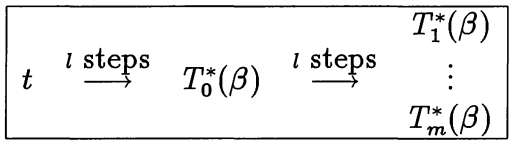

Figure 1: The Markov chain scheme to generate samples under β.

Start by running the scheme l steps to obtain the baseline simulation $T_0^*(\beta)$. Then generate m independent replications $T_1^*(\beta), \ldots, T_m^*(\beta)$ from $T_0^*(\beta)$ through additional independent parallel runs of l steps each. The collection $T, T_1^*(\beta), \ldots, T_m^*(\beta)$ is then exchangeable. This must be so since they are all conditionally independent, with the same distribution, when $T_0^*(\beta)$ is treated as given. Note that the reversibility is crucial. Otherwise the paths to the left and right from $T_0^*(\beta)$ in Figure 1 would lead to different distributions.

The preceding argument was enough for Monte Carlo testing. For exact confidence intervals, we have to impose the additional condition that each sampled $T_j^*(\beta)$ is increasing in β almost surely. One way to achieve this while avoiding computation of the reference density f_0 is to let

$$q_\beta(t'|t) = f_\delta(t')\min\{1, \exp\{(\beta - \delta)(t' - t)\}\}, \tag{5.9}$$

where δ is chosen by the user to enhance efficiency. As it stands, (5.9) is not a proper transition probability function since $\int q(t'|t)dt' < \int f_\delta(t')dt' = 1$, but this can be corrected by redefining $q_\beta(t|t)$ to some positive probability. Imagine that this has been done. From (5.1), and (5.9)

$$f_\beta(t)q_\beta(t'|t) = c(\beta)c(\delta)f_0(t)f_0(t')\min\{\exp(\beta t + \delta t'), \exp(\beta t' + \delta t)\} = f_\beta(t')q_\beta(t|t').$$

This proves that (5.9) corresponds to a reversible Markov process with the desired density f_β as stationary density.

		$l = 5$		$l = 10$		$l = 20$	
α	δ	mean	stan	mean	stan	mean	stan
0.025	0	2.58	0.81	2.22	0.35	2.08	0.20
	1	2.00	0.12	1.98	0.10	1.98	0.09
	2	1.97	0.08	1.97	0.09	1.97	0.09
0.05	0	2.06	0.55	1.79	0.23	1.70	0.14
	1	1.66	0.08	1.66	0.07	1.65	0.07
	2	1.63	0.11	1.65	0.07	1.65	0.06

Table 4: Upper confidence points for the standard normal, observed at $t = 0$, under variation of δ, l, and α. The Monte Carlo sample size was $m = 999$, and the mean and standard deviation are based on 100 runs.

(5.9) can be regarded as a special case of the asymmetrical Metropolis algorithm due to Hastings (1970). Its interpretation is that of a candidate t' being drawn from f_δ and accepted if some uniform random number is less than the second factor on the right in (5.9). If all these candidates and uniforms are stored in advance and kept fixed, the scheme is certain to return values that increase with β, see 5.6 for the proof.

5.4 Numerical illustrations

The purpose of this subsection is to investigate how the performance of the methodology depends on the selection of δ. Two experiments were carried out according to the scheme in Figure 1. The first example assumes that T is normal (β,1) and that $t = 0$ has been observed. Table 4 displays mean and standard deviations of confidence points based on 100 replications for each value of l and δ. Recall that the confidence points all have exact level (in principle). The impact of the Monte Carlo variability depends on how fast the Markov chain forgets the initial state t. This is, in turn, heavily dependent on δ, as is evident from the results in Table 4 (for example, $l = 20$ is not enough when $\delta = 0$). In practice, the speed of convergence will be frightfully slow for most choices of δ, but selections 1-2 standard deviations to the right (left) of the point estimate seem to work fine for upper (lower) confidence points. These recommendations are expected to apply generally. Whether such a data-dependent specification of δ is really consistent with the claim that the confidence points are exact, is discussed in the next section.

The second example comes from logistic regression. Methods for exact inference have been developed by Hirji et al (1987) and Vollset et al (1991) in special cases. The data are taken from the former. There were 46 cases and three covariates, all of the binary type. The null distribution of the interest statistic T, when conditioned on the three other sufficient statistics varied from 19 to 25 according to the distribution shown,

19	20	21	22	23	24	25
0.034	0.183	0.336	0.300	0.123	0.022	0.001

		$l=2$		$l=5$		$l=10$	
α	m	mean	stan	mean	stan	mean	stan
0.025	119	0.216	0.32	0.201	0.25	0.191	0.25
	999	0.135	0.09	0.127	0.08	0.129	0.08
0.05	119	-0.248	0.37	-0.167	0.22	-0.159	0.20
	999	-0.274	0.21	-0.172	0.07	-0.168	0.06

Table 5: Upper critical limits for the slope β of the logistic regression for various combinations of m, α and l. The mean and standard deviations are based on 100 runs.

which was obtained from 10000 simulations. The value actually observed for t was 19, coinciding with the minimum value. Hence only upper limits on β are available in a conditional analysis. Table 5 shows critical points obtained from 100 replications, varying m, α, and l. We took $\delta = 0$ so that in (5.9) $g = f_0$ and $w \equiv 1$. The results do not include randomisation for ties, which in this extreme case brought in an unsatisfactorily large element of unstability. The results in Table 5 seem to become stable from $l = 5$. Note that the upward bias present at $m = 119$, $\alpha = 0.025$ is wiped out at $\alpha = 0.05$.

5.5 Discussion

Much of the flexibility in the basic Metropolis algorithm disappears when the special demands of the paper are imposed. The need to avoid computation of the reference density f_0 is especially restrictive. We have seen that the speed of convergence is sensitive to the choice of δ, and this dependency makes the exactness claimed for the scheme in 5.3 somewhat illusory. If δ is to be selected in a rather narrow interval, it may be hard to argue that it is not data-dependent. If so, the confidence level is no longer exact, although it still holds approximately as $l \to \infty$, since the simulated sample $T_1^*(\beta), \ldots, T_m^*(\beta)$ then becomes free from the influence of the starting position t. The results above suggest that the convergence is quite rapid if δ is selected judiciously. If we take the view that δ in reality depends on the data, it is not obvious that we should use the scheme in Figure 1 with a common root $T_0^*(\theta)$. An alternative is to run m branches in parallel directly from t.

5.6 Proofs

Two statements in this section have remained unproven. Firstly, it must be verifed that (5.8) is really a simulation from (5.1). In fact, this is only true in the limit as $\nu \to \infty$. A heuristic argument runs as follows. The combined probability of $t_1, \ldots, t_\nu$ being produced as the sample from f_δ at the first step *and* t_i being returned by (5.8) at the second step is

$$\frac{\exp\{(\beta-\delta)t_i\}}{\sum_j \exp\{(\beta-\delta)t_j\}} \prod_{j=1}^{\nu} f_\delta(t_j) = c(\delta)\exp(\beta t_i) f_0(t_i) \frac{\prod_{j\neq i} f_\delta(t_j)}{\sum_j \exp\{(\beta-\delta)t_j\}}$$

after inserting for $f_\delta(t_i)$. Integrate this over all t_j except the i'th. Then, by some straightforward algebra

$$P(T^* = t_i) = c_\nu(t_i)\exp(\beta t_i) f_0(t_i),$$

where

$$c_\nu(t_i) = c(\delta)E\{\frac{1}{\nu}\exp(\beta t_i) + \frac{1}{\nu}\sum_{j\neq i}\exp\{(\beta-\delta)T_j\}\}^{-1}.$$

Here the expectation is with respect to all the random variables T_j generated by f_δ, except the one actually picked at the second step. By the law of large numbers

$$c_\nu(t_i) \to c(\delta)\{E\exp\{(\beta-\delta)T_1\}\}^{-1} = c(\beta)$$

as $\nu \to \infty$, and so T^* has the right limiting distribution.

The second statement to prove is that the scheme in 5.3 possesses the required monotonicity with respect to β. Imagine a pool of candidates, say t_c, stored in advance along with uniform random numbers U^*. The algorithm then proceeds according to the recursion

$$t' = \begin{cases} t_c & \text{if } U^* \leq \exp\{(\beta-\delta)(t_c - t)\} \\ t & \text{otherwise .} \end{cases} \tag{5.10}$$

We claim that the current t at each step will be at least as large under β_1 as under β if $\beta_1 > \beta$. Note that the start of the scheme is the same for any β. Suppose that t_1 and t, with $t_1 \geq t$, are the values under β_1 and β at a certain stage in the iteration. We have to show that $t_1' \geq t'$ one step ahead. The reasoning depends on where the candidate for replacement, the same one, t_c, for both schemes, is located. (i) If $t \leq t_c \leq t_1$, then, from (5.10), $t_1' \geq \min(t_c, t_1) = t_c = \max(t_c, t) \geq t'$. (ii) Suppose $t_c > t_1$. There is nothing to prove if $t' = t$ (since $t_1 \geq t$), but suppose $t' = t_c$. Then, by (5.10), $U^* \leq h(\beta)$, where $h(\beta) = \min\{1, \exp\{(\beta-\delta)(t_c - t)\}\}$. Similarly, let $h_1(\beta) = \min\{1, \exp\{(\beta-\delta)(t_c - t_1)\}\}$ and note, by the assumptions made, $h(\beta) \leq h(\beta_1) \leq h_1(\beta_1)$ so that $U^* \leq h_1(\beta_1)$. Hence $t_1' = t_c$, and the β_1-scheme has not fallen below the other. (iii) The remaining possibility $t_c < t$ is handled as in (ii), except for the inequalities being reversed.

6 Discussion and further work

We have made the point that the traditional construction of confidence estimates from tests of significance may be applied to Monte Carlo tests. Potential benefits are the enormous versatility of this class of tests and their high level accuracy. Research is needed on several fronts. The present paper has only dealt with strictly pivotal statistics that made the confidence levels exact. Moreover, to avoid complicated and uninteresting sets as solutions, we placed a severe restriction on the test statistics by forcing the sampled realisations to be monotone (almost surely) in the parameter of interest. In applications this has demanded considerable inventiveness. Different representations and algorithms were necessary each

time. It is not known how often the set-up in 2.2 can be employed in practice, but it is tempting to look to the large body of sampling techniques available.

Wider applicability is reached if the monotonicity condition is thrown away at the cost of exactness. A general *conservative* method was suggested in (2.5). The level accuracy of this proposal has not been studied, although we have expressed hope that it will remain high. There is an asymptotic argument in support of this. When the number of observations grows, many statistics behave more and more like the location type, which permits, as we have seen, exact interval estimates to be constructed. It might be worthwhile to study the level error of the conservative method asymptotically and numerically.

Asymptotics is also likely to be useful when the model contains unknown quantities other than the parameter of interest. These nuisance parameters (or distributions) must be estimated prior to the simulation. The Monte Carlo tests are no longer exact, although their level errors are known to be small. By the reasoning in section 2, the confidence set C in (2.4) inherits the level accuracy from its parent Monte Carlo test. This motivates research into how unknown quantities should be integrated in the methodology. There are several issues. The small level errors Hall and Titterington (1989) found for Monte Carlo tests were due to the test statistics being asymptotically pivotal. It should be investigated how this property can be combined with the other demands we have been making. Another point arises with the control exercised over the estimates that are used with the sampling routine. This creates possibilities. Bølviken and Skovlund (1994) argue that the level errors of Monte Carlo tests can be brought further down by adjusting these estimates in the right way. The same idea could work with confidence estimation.

Monte Carlo confidence intervals, as envisioned, in principle compete with those found in the bootstrap literature. Our approach can hardly match the enormous generality and flexiblity of the bootstrap/resampling methods and neither the automated way these methods can be implemented. It could be, however, that by working from stronger assumptions, more accurate methods can be found in special cases. We believe many small-sample situations to be well served by the ideas of the present paper.

References

Babu, G.J., and Rao, C.R. (1993), Bootstrap methodology, in *Computational statistics,* Handbook of statistics, vol. 9, Ed. C.R. Rao, New York: North-Holland.

Barnard, G. (1963), Contribution to the discussion of Bartlett's paper, *Journal of the Royal Statistical Society*, Ser. B, 25, 294.

Beran, R. (1988), Balanced simultaneous confidence sets, *Journal of the American Statistical Association*, 83, 679-686.

Beran, R. (1988), Prepivoting the test statistics: A bootstrap view of asymptotic refinements, *Journal of the American Statistical Association*, 83, 687-697.

Besag, J., and Clifford, P. (1989), Generalized Monte Carlo significance tests, *Biometrika*, 76, 633-642.

Boswell, M.T., Gore, S.D., Patil, G.P., and Taille, C. (1993), The art of computer generation of random variables, in *Computational statistics,* Handbook of statistics, vol. 9, Ed. C.R. Rao, New York: North-Holland.

Bølviken, E., and Skovlund, E. (1994), Improving the level error of Monte Carlo tests, *in preparation.*

Devroye, L. (1986), *Non-uniform random variate generation,* New York: Springer-Verlag.

Efron, B., and Tibshirani, R.J. (1993), An introduction to the bootstrap, New York: Chapman and Hall.

Garthwaite, P.H., and Buckland, S.T. (1992), Generating Monte Carlo confidence intervals by the Robbins-Monro process, *Applied Statistics*, 41, 159-171.

Gelfand, A.E., and Smith, A.F.M. (1990), Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, 85, 398-409.

Hall, P. (1988), Theoretical comparison of bootstrap confidence intervals, *The Annals of Statistics*, 16, 927-953.

Hall, P., and Titterington, D. M. (1989), The effect of simulation order on level accuracy and power of Monte Carlo tests, *Journal of the Royal Statistical Society*, Ser.B, 51, 459-467.

Hastings, W.K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, 57, 97-109.

Hirji, K.F., Metha, C.R., and Patel, N.R. (1987), Computing distributions for exact logistic regression, *Journal of the American Statistical Association*, 82, 1110-1117.

Hollander, M., and Wolfe, D.A. (1973), *Nonparametric statistical methods*, New York: Wiley.

Jöckel, K.-H. (1986), Finite sample properties and asymptotic efficiency of Monte Carlo tests, *The Annals of Statistics*, 14, 336-347.

Lai, T.L., and Robbins, H. (1979), Adaptive design and stochastic approximation. *The Annals of Statistics*, 7, 1196-1221.

Lehmann, E.L. (1986), *Testing statistical hypotheses,* second edition, New York: Wiley.

Neal, R. (1993), "Probabilistic inference using Markov chain Monte Carlo methods", Technical Report, Department of Computer Science, University of Toronto.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1986), *Numerical Recipes. The art of scientific computing,* Cambridge: Cambridge University Press.

Ripley, B. (1987), *Stochastic simulation,* New York: Wiley.

Vollset, S.E., Hirji, K.F., and Elashoff, R.M. (1991), Fast computation of exact confidence limits for the common odds ratio in a series of 2x2 tables, *Journal of the American Statistical Association*, 86, 404-409.